ReBaIR: Reference-Based Image Restoration

Michael Bernasconi[1,2] Abdelaziz Djelouah[2] Yang Zhang[2] Markus Gross[1,2] Christopher Schroers[2]
[1]ETH Zürich [2]DisneyResearch|Studios
michael.bernasconi@inf.ethz.ch abdelaziz.djelouah@disney.com

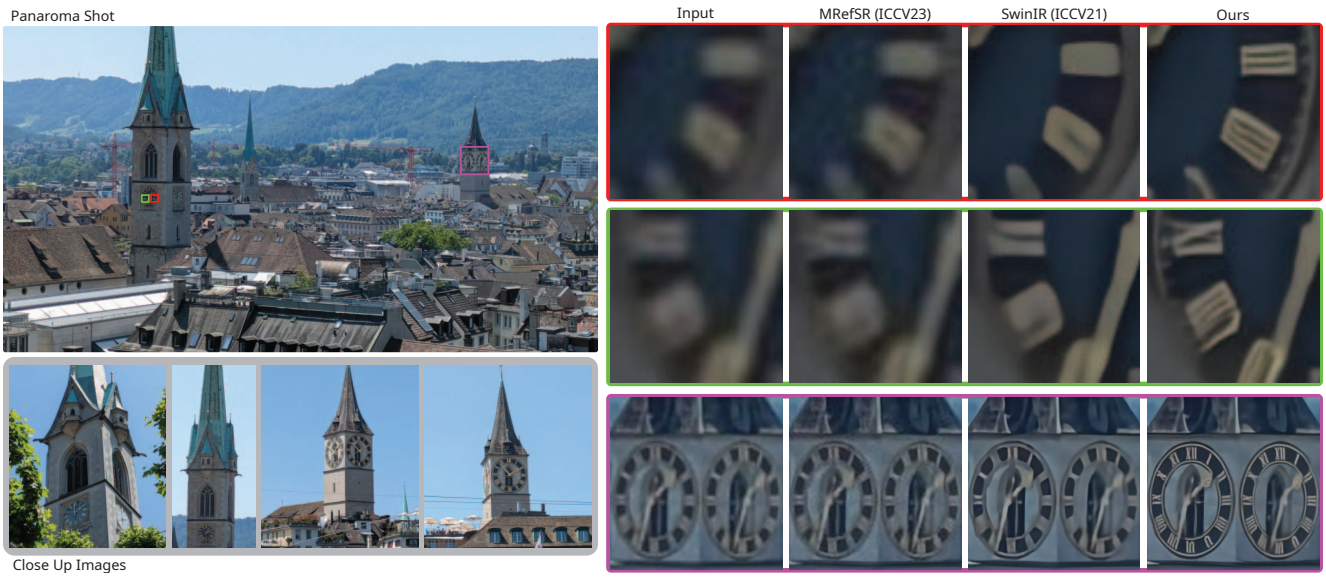

Figure 1. Visual comparison between our method, MRefSR [25], and SwinIR [13] on a real-world example where a panorama shot is upscaled with the help of a set of close up images captured during a trip through the city. All images were captured using a Google Pixel 7a smartphone. Our method uses SwinIR trained for the RealSR setting as a backbone. While MRefSR struggles in this in-the-wild scenario, our method effectively uses the available reference images and produces high quality results.

Abstract

In recent years image restoration methods have made significant progress addressing a wide variety of degradations. Some methods focus on a specific task while others address enhancement in a more generic setting. Reference-based restoration aims at leveraging any image of higher quality that would be available to further improve the results. To the best of our knowledge reference images were only used in the context of image and video super-resolution (RefSR and RefVSR), with specialized models. In this work, we propose a novel and generic reference-based restoration method that is applicable to any model and any task. We start with the observation that restoration models typically operate in feature space before a final decoding step which transforms the extracted features into an image. Our model operates as an add-on that extracts information from the references and uses this to enhance these pre-decoding feature maps, leading to significant improvement in image quality. Our strategy is compatible with virtually all existing image restoration methods and we demonstrate this with a wide range of both specialized and generic image enhancement models where we achieve a significant boost in quality. Besides its generic applicability, we also demonstrate that the proposed solution outperforms existing specialized RefSR state-of-the-art methods both quantitatively and qualitatively.

1. Introduction

Image quality enhancement is a fundamental computer vision task which aims to restore a high quality signal from a low quality image. As image quality can be degraded

for a multitude of reasons there is a variety of methods aimed at different sub-tasks of image restoration. To name a few, there are specialized methods for image super-resolution [5, 7–9, 13, 18, 26], compression artifact removal [13], and denoising [13, 22]. In each of these categories significant progress has been made in recent years and current state-of-the-art methods are able to recover impressive levels of detail from heavily degraded input images.

In many real-world applications a single degraded image is not the only information available. Often, higher quality images depicting similar content are available and can be used to guide the restoration process to achieve even higher quality. In the field of image super-resolution a number of methods have been proposed which are able to utilize available reference images and show a significant boost in quality. However existing reference based super-resolution (RefSR) methods [2, 10, 14, 19, 21, 24, 25, 27, 29] are specifically designed and trained (from scratch) for the task of super-resolution. We believe this to be an important limitation. On one side it is difficult to benefit from progress made in terms of models (architecture, training, etc.), as each time a new design and re-training is needed to integrate reference image information. On another side, there is no reason to limit the usage of references to the super-resolution task.

In this work we propose a generic reference refinement module that is designed as an add-on compatible with any existing image restoration method, enabling the optional use of reference images to boost quality. By extracting information from the references and using it to enhance the last feature maps of any existing restoration model, our proposed module can be efficiently trained as we leave the original model frozen, while benefiting from all its advantages. Compared to existing RefSR approaches this has two main advantages: First, our approach allows us to build upon the large existing body of work exploring neural network architecture and training for super-resolution, and we can easily utilize newer, more powerful image models once they become available. Second, our generic design is not limited to the task of RefSR. Instead our method can be applied to any image restoration task. We show that our approach yields state-of-the-art results on the task of RefSR and can be applied seamlessly to other image restoration tasks like denoising or compression artifact removal.

We claim the following contributions:

- A novel reference-based restoration module compatible with existing image restoration methods.
- We achieve state-of-the-art quantitative and qualitative results for the task of RefSR.
- The first method capable of utilizing reference images for general image restoration tasks like denoising or compression artifact removal.

2. Related Work

Image restoration is a classic computer vision task with the goal of recovering a high-quality signal from a degraded, low-quality image. Dong *et al.* [7] were among the first to apply deep neural networks to the task of image super-resolution, which have since become the standard approach of tackling any image restoration task. Since then, significant progress in network architecture [8, 13, 18, 26] has been made and current state-of-the-art image restoration methods are able to recover impressive details even from heavily degraded inputs. Methods like BSRGAN [23] are even able to upscale low-resolution images with a multitude of degradations while methods like [1, 5, 9] can handle arbitrary scaling factors or even arbitrary geometric transformations [1, 16]. In an effort to further increase image quality methods have been proposed which are able to utilize additional input information. For example, video super-resolution methods like [3, 4, 20] utilize temporal information from neighboring video frames to produce high quality super-resolution results, while methods like [1, 6] conditioning their method on the downsampling kernel. Another popular approach to increase super-resolution quality is the usage of reference images. Zheng *et al.* [28] were among the first to propose a method of this task. Later methods such as [11, 14, 19, 21, 27] improved upon these early results by identifying better correspondences between the degraded and reference image. Other methods such as [2, 10, 24] have further pushed quality by employing more sophisticated architectures and training procedures. A major limitation of most RefSR methods is their inability to utilize multiple reference images. This was addressed by MRefSR [25] which was the first RefSR method capable of utilizing multiple reference images. In addition to their method they also introduce the LMR dataset containing both training and testing examples for multi-RefSR. While their method clearly outperforms previous methods which were limited to a single reference image our evaluation shows that our two stage refinement process manages to extract significantly more information from the available reference images and produces much higher quality results. The task of RefSR has also been extended to video by RefVSR [12]. Their method is, however, limited to their highly specific triple camera setup and only a single reference image can be used for each video frame. To the best of our knowledge currently no reference-based methods exist which can handle image restoration tasks other than super-resolution or even arbitrary scale super-resolution. Additionally, all RefSR methods so far have been designed and trained from scratch for RefSR. This means integrating new advances in image restoration into these existing approaches requires redesign their architecture and computationally expensive re-training. Our method on the other hand can easily be adapted to any new backbone without

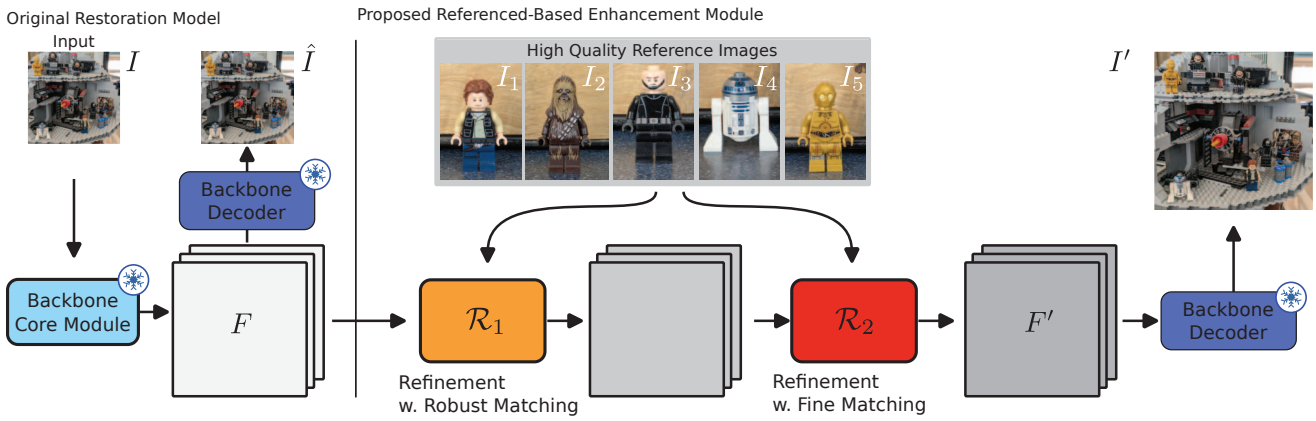

Figure 2. Overview of our proposed reference-based image restoration method. First, an initial feature map is extracted from the degraded input image using the pre-trained backbones feature extractor. Both our proposed refinement stages then sequentially refine this feature map by injecting information from the available high quality reference images. Finally, the refined feature map is decoded to an image using the backbones image decoder. Note that only the two refinement stages are trained while all backbone parts are kept frozen.

any redesign and relatively low computational cost.

3. Method

On a high level all existing image restoration methods operate in two stages as illustrated on the left side of Figure 2. Starting from the input image I, the core part of most existing enhancement models operates on features, the resulting features F are then decoded into the enhanced image $\hat{I}$. In the following we assume that F has the same spatial resolution as $\hat{I}$. Typically for SR methods this means that F is the high resolution feature map.

We propose a reference-based image restoration module that leaves the backbone architecture unchanged. Using a set of available reference images $I_{1,\cdots,N}$, our method refines F in two stages. Both follow the same principles: first, compute a mapping between the original and the references; second, extract and align features from the references; third, use the aligned features to enhance F. The first stage relies on robust matching to compute the mapping with the references, while the second stage relies on a fine matching methods. The motivation for these two stages is the ability to best leverage the references with both large and small differences in appearance and viewpoint.

In summary, starting from the backbone core module $\mathcal{E}$, we enhance the initial feature map F through the two stages R_1 and R_2 using the references, before finally decoding the result image I' with the backbones decoder $\mathcal{D}$.

$$F = \mathcal{E}(I) \tag{1}$$
$$F' = \mathcal{R}_2(\mathcal{R}_1(F)) \tag{2}$$
$$I' = \mathcal{D}(F') \tag{3}$$

For the robust matching we rely on the same strategy as RefSR, using CCN [11] based matching, while for the fine matching we use PDCNet+ [17].

3.1. Aligned Multi-Scale Feature Extraction (A-MSFE)

Reference image features are extracted at multiple scales. Feature at each scale are used to progressively enhance the feature map F extracted from the backbone core module in a coarse-to-fine manner. We start by describing the multi-scale architecture then detail the spatial feature alignment. Here we assume the $2D$ mappings (M_i) to the references I_i is already computed.

Multi-Scale Feature Extraction (MSFE). The multi-scale feature extractor (MSFE) is the basic building block used throughout our method. From any feature map F it returns three feature maps F^1, F^2, F^3 at scales 1, $\frac{1}{2}$, and $\frac{1}{4}$ respectively.

$$F^l = \text{RDB}(\text{down}(F^{l-1})) \quad \text{and} \quad F^1 = \text{RDB}(F) \tag{4}$$

Here, RDB refers to residual-dense-blocks as introduced in [18, 26] and they are used to extract features at each scale. The *down* operation spatially downscales feature maps by a factor of 2 using pixel-unshuffling followed by a convolution adjusting the number of channels.

Pre-Alignment MSFE. When using pre-alignment the multi-scale feature extraction is only used after $2D$ alignment

$$F_i = \text{RDB}(\text{Conv}_{\text{in}}(I_i)) \tag{5}$$
$$\tilde{F}_i = \text{warp}(F_i, M_i) \tag{6}$$
$$\tilde{F}_i^1, \tilde{F}_i^2, \tilde{F}_i^3 = \text{MSFE}(\tilde{F}_i). \tag{7}$$

To avoid warping the raw reference image I_i directly we perform shallow feature extraction via RDB first which

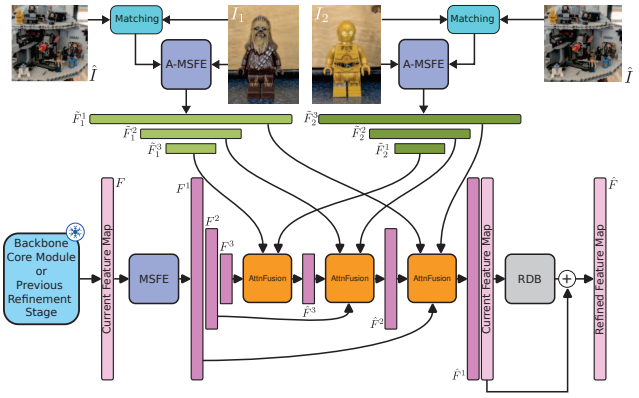

Figure 3. Visual illustration of our proposed multi-scale multi-reference feature refinement module. The current feature map is refined at multiple scales by the MRFR module which aggregates and injects information from multiple reference images. Note that in the first refinement stage the *current feature map* is the feature map provided by the backbone while in the second refinement stage it refers to the output of the first refinement stage.

results in the unaligned feature map F_i. This feature map F_i is then warped and further processed by the MSFE.

Post-Alignment MSFE. When using post-alignment the reference image I_i is first processed by the MSFE and each extracted feature map is then aligned to F individually.

$$F_i^1, F_i^2, F_i^3 = \text{MSFE}(\text{Conv}_{\text{in}}(I_i)) \tag{8}$$

$$\tilde{F}_i^l = \text{warp}(\tilde{F}_i^l, M_i^l) \tag{9}$$

Each scale l is warped according to the mapping M_i adjusted for the scale. Further details regarding how the mappings are obtained are discussed next.

We note that multi-scale feature extraction is possible with *Pre-* or *Post-* alignment. Experimentally their performance varies depending on the type of matching we use (robust *vs* fine). We use post-alignment MSFE for the enhancement with robust matching and pre-alignment MSFE for the enhancement with fine matching.

3.2. Matching and Warping

Our aligned multi-scale feature extraction relies on the estimation of a $2D$ mapping between the image to enhance and the references. There can be a wide range of variability between the references and the image to enhance, both in terms of viewpoint and colors. To handle more extreme cases while still benefiting from refined matching when possible, we adopt a two-stage strategy.

Robust Matching. Following MRefSR [25] we use the Contrastive Correspondence Network (CCN) introduced by C^2-Matching [11] to extract correspondences between the base image $\hat{I}$ and each reference image I_i. We refer to this mapping as M_i^3. Note that M_i^3 is at one fourth the resolution of $\hat{I}$. We generate higher resolution version of this mapping M_i^2 and M_i^1 at half and full resolution via nearest neighbor upsampling. The main advantage of CCN is its ability to identify similar content in reference images for each location in the base image even under large deformations.

Precise Matching. For our second refinement stage we use PDCNet+ [17] to extract correspondences between the image $\hat{I}$ and each reference image I_i. The main advantage of PDCNet+ is that it provides us a smooth flow field at full resolution with sub-pixel accuracy. Warping is performed using nearest neighbor grid sampling and adding the sub-pixel offsets and confidence provided by PDCNet+ to the warped feature maps along the channel dimension. More details are provided in supplementary material.

3.3. Multi-Scale Multi-Reference Feature Refinement

Our proposed multi-reference feature refinement is illustrated in Figure 3. First, we operate on the core feature map F. Here we use MSFE to extract features at three different scales, which are progressively enhanced using the references. On the reference side (top part of the figure), the mapping estimated between $\hat{I}$ and the reference I_i is used to extract multi-scale features using the aligned multi-scale feature extraction. We now describe in more details how we produce the refined feature map.

Given the multi-scale feature extracted from F

$$F^1, F^2, F^3 = \text{MSFE}(F), \tag{10}$$

and the aligned multi-scale features extracted for each reference

$$\tilde{F}_i^1, \tilde{F}_i^2, \tilde{F}_i^3 = \text{A-MSFE}(I_i)\ \forall i \in [1, N], \tag{11}$$

we use an attention based fusion mechanism (**AttnFusion**) to fuse the information at each scale, before propagation to the next scale.

$$\hat{F}^3 = \text{AttnFusion}(F^3, \tilde{F}_1^3, \cdots, \tilde{F}_N^3) \tag{12}$$

$$\hat{F}^l = \text{AttnFusion}(F^l + \text{up}(\hat{F}^{l+1}), \tilde{F}_1^l, \cdots, \tilde{F}_N^l) \tag{13}$$

where **AttnFusion** is based on multi-head attention (mha)

$$F_{\text{ref}} = \text{mha}(F, \tilde{F}_i, \cdots, \tilde{F}_N)$$

$$\hat{F} = F + \text{Conv}_{\text{out}}(\text{RDB}(F||F_{\text{ref}}))$$

F is used to extract queries, while key-value pairs are extracted from $F_{1,\cdots,N}$. The attention mechanism is performed separately for each spatial location. Note that the operator $||$ refers to concatenation along the channel dimension. The obtained feature map is used as residual to F (See Figure 3).

Method	LMR			CUFED5			WR-SR		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
LIIF [5] (CVPR21)	28.63	0.835	0.157	25.78	0.779	0.193	27.44	0.796	**0.209**
SwinIR [13] (ICCV21)	29.87	0.859	0.183	26.92	0.812	0.199	28.32	0.817	0.245
DRCT [8] (CVPR24)	30.00	0.861	0.183	27.02	0.812	0.205	28.40	0.818	0.246
C^2-Matching-*rec* [11] (CVPR21)	30.52	0.880	0.147	28.18	0.852	0.137	28.19	0.814	0.244
DATSR-*rec* [2] (ECCV22)	30.89	0.888	0.132	28.58	0.863	0.122	28.19	0.815	0.234
MRefSR-*rec* [25] (ICCV23)	31.78	0.903	0.121	28.80	0.868	0.126	28.40	0.818	0.237
Ours (LIIF)	32.18	0.912	0.105	28.94	0.874	0.116	28.55	0.823	0.230
Ours (SwinIR)	32.61	**0.918**	0.100	29.29	0.882	**0.108**	28.75	0.829	0.224
Ours (DRCT)	**32.68**	**0.918**	**0.098**	**29.33**	**0.883**	0.110	**28.80**	**0.830**	0.222

Table 1. Numeric evaluation for $\times 4$ scaling on common RefSR datasets. We highlight the **best** result in bold and underline the second-best. Our method using either SwinIR or DRCT as a backbone significantly outperforms existing RefSR approaches. Even when LIIF is used as a backbone our method outperforms existing approaches while not being limited to fixed $\times 4$ scaling.

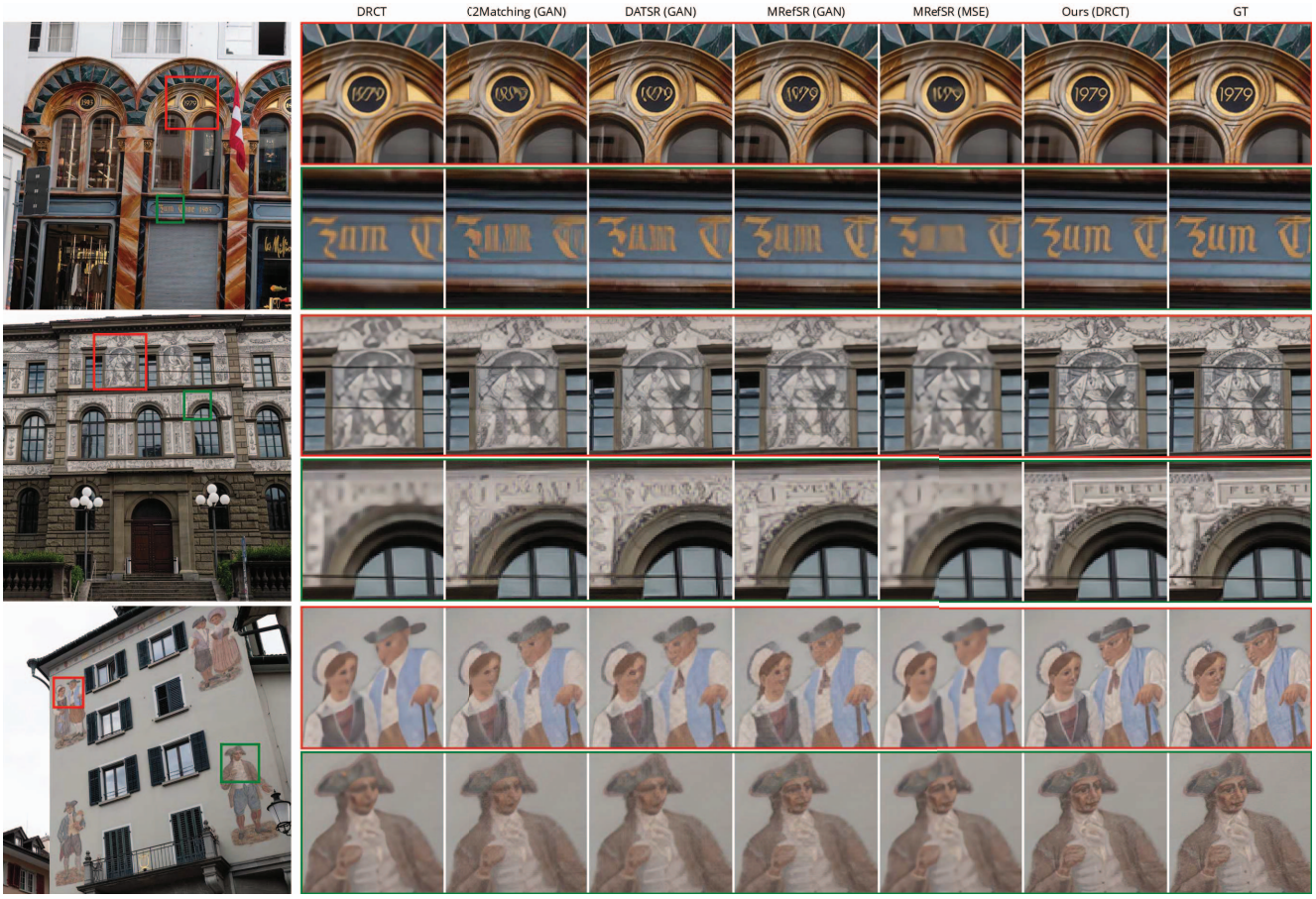

Figure 4. Visual results for $\times 4$ scaling. Previous methods either produce noticeable artifacts when trained with an adversarial loss (column 2-4) or blurry results when trained without (column 5). Our method utilized reference images more effectively managing to produce sharp results without any noticeable artifacts.

4. Results

We evaluate our method on a variety of image restoration tasks using a variety of backbone models. For each experiment our model is trained in two stages. First, the refinement stages 1 and 2 are trained independently, each for 300k steps. Both stages are then combined and trained jointly

for another 150k steps. All training is performed on the LMR [25] dataset using two RTX4090 GPUs. Further training details are provided in supplemental material.

4.1. Ref-SR

We start our evaluation by comparing our method to existing RefSR approaches. Note that previous methods train two versions of their model. A perceptually optimized version trained with an adversarial loss and a numerically optimized version without adversarial loss. We indicate the numerically optimized versions with the postfix *-rec*. For our method we use a single version trained with $L1$ loss for both numeric and visual evaluation. Also note that DATSR [2] and C^2-Matching [11] are both limited to using only a single reference images. If multiple reference images are available we simulate a best case scenario for these methods by picking the reference image that results in highest PSNR. In Table 1 we present a numeric evaluation for $\times 4$ reference based super-resolution on common datasets. For this evaluation we train our method using both SwinIR [13] and DRCT [8] as a backbone. With either backbone our method clearly outperforms existing approaches across the board. Even when using LIIF [5], an arbitrary scaling SR method, as a backbone our method still produces state-of-the-art results. When using the LIIF backbone our method is not fixed to $\times 4$ scaling but can handle arbitrary scaling factors instead. Our methods strong numeric performance is backed up by clear visual improvements which is illustrated in Figure 4. Here, we see that previous methods trained with adversarial loss (GAN) produce undesirable artifacts while MSE optimized methods produce overly blurry results. Our method produces sharp results without any noticeable artifacts.

4.2. Ref-Restoration

As our method is naturally compatible with most existing image restoration methods we evaluate the benefits of reference images on a variety of image restoration tasks. Table 2 illustrates that our method manages to effectively use reference images to improve quality for arbitrary scaling SR using LIIF [5] as a backbone, JPEG artifact removal using SwinIR [13] as a backbone, denoising using Restormer [22] as a backbone, and RealSR using SwinIR [13] as a backbone. For each task we use the publicly available pretrained backbone checkpoints. Figures 5&6 illustrate the visual improvement our method achieves for different restoration tasks. We can clearly see that our method effectively uses the available reference images to improve quality in each case. Our methods benefits also translate to the in-the-wild setting presented in Figure 1. Here, we upscale a panorama shot captured with a smartphone using a set of close up images. The set of close up reference images was captured using the same smartphone during a walk through the city. Compared to MRefSR [25] our method produces much sharper, more visually appealing results.

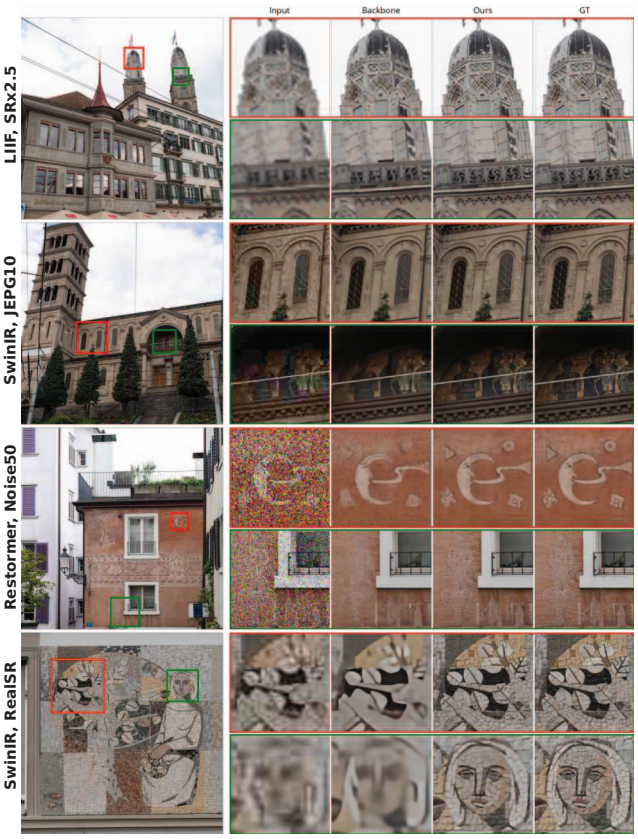

Figure 5. Visual results illustrating the benefits of using reference images for a variety of tasks. Our method effectively utilized available reference images which significantly boost quality compared to the backbone.

Task	Backbone	Backbone			Backbone + Ours		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
SR $\times 2.5$	LIIF	29.14	0.866	0.076	31.10	0.892	0.064
SR $\times 6.5$	LIIF	25.70	0.719	0.258	26.41	0.750	0.266
JPEG 10	SwinIR	32.16	0.895	0.142	33.69	0.924	0.101
JPEG 30	SwinIR	36.18	0.951	0.065	37.21	0.960	0.052
Noise 25	Restormer	34.39	0.937	0.057	36.62	0.954	0.048
Noise 50	Restormer	29.40	0.870	0.108	33.78	0.926	0.083
RealSR	SwinIR	25.62	0.719	0.347	27.53	0.794	0.239

Table 2. Numeric evaluation illustrating the benefits of using reference images for different tasks. We evaluate arbitrary scale super-resolution for scales 2.5 and 6.5 using LIIF as backbone, JPEG artifact removal for quality levels 10 and 30 using the SwinIR backbone, image denoising with the Restormer backbone and σ 25 and 50, and finally real image super-resolution using SwinIR as a backbone. For RealSR we apply noise to the HR image with $\sigma = 20$, then downscale the image by $\times 4$, and finally apply JPEG compression with $q = 30$ to the LR image.

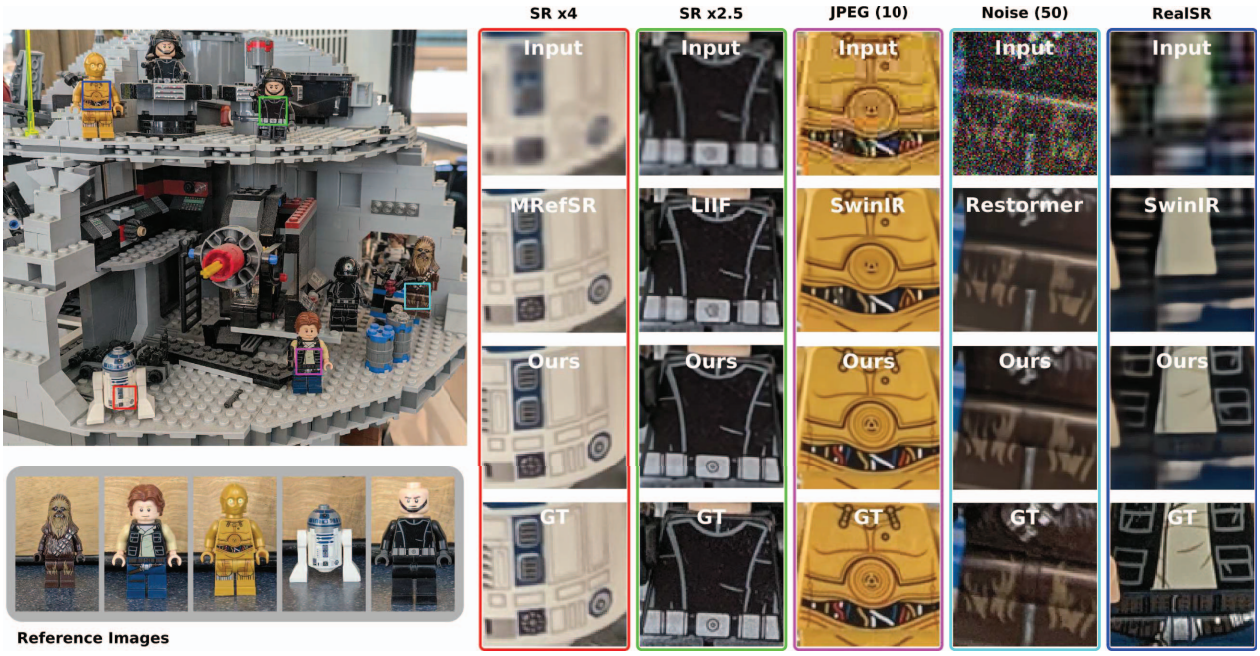

Figure 6. Visual results of our method applied to a variety of tasks. For $SR \times 4$ we compare our method against MRefSR [25] which is also capable of using the available reference images. For all other tasks we compare our method against its backbone. For the task of RealSR the input image is generated by first adding noise ($\sigma = 20$) to the ground truth image, then downscaling by a factor of 4, and finally applying JPEG compression ($q = 30$) to the LR image.

4.3. Ablation Study

In this section we ablate a number of design choices we made throughout our method. As existing reference-based methods focus on the task of super-resolution we choose this task to ablate our design choices. However, the insights gained from our ablation study should translate to other tasks and our evaluation shows that our method performs well on a wide variety of reference-based restoration tasks. Throughout this section we report runtime, GPU memory consumption, and number of parameters in different settings. Unless mentioned explicitly results are generated on the LMR [25] testset without performing any CPU offloading and transfer times from CPU to GPU are not included in the reported runtimes. Runtime is always reported as seconds-per-image which refers to the average number of seconds required to process each example in the testset. Memory consumption refers to maximum GPU memory consumption over the whole testset in Gigabytes. The number of parameters is reported in three categories. *Train* refers to the number of parameters which are optimized when training a given method. *Enhance* refers to the number of parameters which are directly contributing to the enhanced output image. This excludes parameters from PDCNet+ [17] and CCN [11] which are only used to compute correspondences between the base and reference images. *Total* refers to the total number of parameters in the model. Unless explicitly stated otherwise our ablation is performed using the $\times 4$ upscaling SwinIR [13] as a backbone.

Corr	Align	Scales	PSNR	SSIM	LPIPS	Time [s]	Mem [GB]	Params [M] Train	Enhance	Total
CCN	Post	3	**31.55**	**0.899**	**0.125**	8.4	8.0	12.5	24.4	25.5
CCN	Pre	3	31.39	0.897	0.126	8.4	8.0	12.5	24.4	25.5
CCN	Post	1	30.96	0.888	0.140	8.6	12.8	3.3	15.2	16.3
PDC	Pre	3	**32.44**	**0.916**	**0.101**	2.6	8.1	12.5	24.4	42.8
PDC	Post	3	32.37	0.914	0.104	2.5	8.1	12.5	24.4	42.8
PDC	Pre	1	32.00	0.909	0.109	2.9	12.9	3.5	15.4	33.8
CCN + PDC		3	**32.61**	**0.918**	**0.100**	10.0	8.1	24.9	36.8	56.4

Table 3. Ablation study showing effect of different model architecture choices in both refinement stages. For the CCN-alignment stage post-aligning feature maps performs better than pre-alignment, while for the PDC-alignment stage pre-alignment performs better. For both stages a deeper 3-scale architecture with $64, 128, 256$ channels produces higher quality results and is less computationally expensive compared to a 1-scale architecture with 128 channels. In the last row we show the result of combining the best performing CCN and PDC stages. The combined model outperforms both individual refinement stages which clearly shows the advantage of our two stage refinement process.

Our first ablation is presented in Table 3. Here, we show that post-alignment performs better when CCN correspondences are used while pre-alignment is preferable for PDCNet+ correspondences. Additionally, we also show that our multi-scale design has clear benefits both in terms of quality and computational cost. We compare a single-scale version of our method where all refinement happens at full resolution with 128 feature channels to our multi-scale de-

sign with three levels of size 64, 128, and 256. The results show that the multi-scale design performs significantly better while using less GPU memory and achieving faster runtime. We note that the PDCNet+ refinement stage performs significantly better at greatly reduced computational cost compared to the CCN stage. This may be an interesting direction to pursue for future work focused on resource constrained environments. Our work is, however, aimed at maximum performance in an offline setting and we see that combining the two refinement stages yields the best results. This illustrates the effectiveness of our two stage refinement process.

Corr	Match Image	Enhancement	PSNR ↑	SSIM ↑	LPIPS ↓
CCN	bicubic	features	31.55	0.899	0.125
CCN	backbone	features	31.44	0.897	0.127
CCN	bicubic	image	31.53	0.899	0.125
PDC	backbone	features	32.44	0.916	0.101
PDC	bicubic	features	32.36	0.914	0.105
PDC	backbone	image	32.40	0.915	0.105

Table 4. Ablation study showing the effect of using different match images and injected reference information at different locations.

Other important design choices are which image to use when computing correspondences, where to inject information from reference images, and the ordering of the two refinement stages. We ablate this in Table 4. For CCN correspondences we find that it is best to use a biubically upscaled image to extract correspondences while for PDCNet+ an already enhanced image from the backbone model performs better. This finding also informs our ordering of the refinement stages. Using CCN for the first stage makes sense as it does not benefit from an already enhanced image. PDCNet+ is best suited for the second stage where it can benefit from an already enhanced image from the first stage. For both stages we see that refining the feature map extracted by the backbone performs better than refining the restored image directly.

In Table 5 we showcase the effect both the quality and number of reference images have on computational cost and performance. This experiment is performed on the CUFED5 [27] dataset which provides a set of reference images with different levels of similarity for each example. We evaluate our method using the best/worst 1/3 reference images and also using all the available reference images. We identify two key insights from these results - more similar reference images are better than less similar ones and more reference images are better than fewer. We also note that using all available reference images yield the best results which indicates that our method manages to ignore information from lower quality reference images where higher quality information is available.

Finally, we compare our method to MRefSR [25] both

References	PSNR	SSIM	LPIPS	Time [s]	Mem [GB]
Best 1	28.84	0.869	0.124	0.3	0.8
Best 3	29.23	0.880	0.110	0.4	1.1
Worst 1	27.30	0.822	0.184	0.3	0.8
Worst 3	28.06	0.848	0.149	0.4	1.1
All	29.29	0.882	0.108	0.6	1.5

Table 5. Impact of number and quality of reference images on our method computational cost and performance on the CUFED5 [27] testset which provides reference images of varying similarity level for each example. We see that higher quality reference images result in better performance compared to lower quality references. We also see that, independent of reference image quality, providing more reference images strictly improves output quality.

Method	PSNR	SSIM	LPIPS	Time [s]	Mem [GB]	Params[M] Train	Enhance	Total
SwinIR	29.87	0.859	0.183	0.8	1.4	11.9	11.9	11.9
DRCT	30.00	0.861	0.183	4.1	3.1	27.6	27.6	27.6
MRefSR (MSE)	31.78	0.903	0.121	13.6	22.1	23.7	23.7	25.4
Ours S (SwinIR)	32.26	0.913	0.107	9.1	5.0	6.1	18.0	37.6
Ours S (DRCT)	32.36	0.914	0.106	12.2	5.2	6.1	33.7	53.3
Ours L (SwinIR)	32.61	0.918	0.100	10.0	8.1	24.9	36.8	56.4
Ours L (DRCT)	32.68	0.918	0.098	13.0	8.3	24.9	52.5	72.1

Table 6. Computational cost and quality comparison between our method and MRefSR [25]. Independent of the backbone our method significantly outperforms MRefSR in both quality and computational cost.

in terms of runtime and peak memory consumption on the LMR [25] dataset. Our method is evaluated using both the SwinIR [13] and DRCT [8] as a backbone and for both options we test a small and large version. The large version (*Ours L*) uses intermediate feature maps of size $64, 128, 256$ while the small version (*Ours S*) halves each layers size to $32, 64, 128$. The results are presented in Table 6. We see that, in all configurations, our method produces higher quality results than MRefSR [25] while consuming less memory and achieving faster runtimes. We also see that the small versions of our method have significantly lowered memory requirements while still providing noticeably higher quality than other methods. While in this work we focus on maximum quality in a offline setting we believe that the results achieved by our small model show that application in a more resource constrained environment is feasible. This is, however, beyond the scope of this work.

5. Conclusion

In this paper we have introduced a generic method capable of utilizing available reference images for any image restoration task. Our method is compatible with virtually any image restoration backbone and our results clearly show the benefits of using reference images across a variety of image restoration tasks. For the task of RefSR our method significantly outperforms the current state-of-the-art while simultaneously reducing the required computational resources.

References

[1] Michael Bernasconi, Abdelaziz Djelouah, Farnood Salehi, Markus Gross, and Christopher Schroers. Kernel aware resampler. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22347–22355, 2023. 2

[2] Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *European conference on computer vision*, pages 325–342. Springer, 2022. 2, 5, 6

[3] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4947–4956, 2021. 2

[4] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022. 2

[5] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 2, 5, 6

[6] Victor Cornillère, Abdelaziz Djelouah, Wang Yifan, Olga Sorkine-Hornung, and Christopher Schroers. Blind image super resolution with spatially variant degradations. *ACM Transactions on Graphics (proceedings of ACM SIGGRAPH ASIA)*, 38(6), 2019. 2

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 2

[8] Chih-Chung Hsu, Chia-Ming Lee, and Yi-Shiuan Chou. Drct: Saving image super-resolution away from information bottleneck. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6133–6142, 2024. 2, 5, 6, 8

[9] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1575–1584, 2019. 2

[10] Yixuan Huang, Xiaoyun Zhang, Yu Fu, Siheng Chen, Ya Zhang, Yan-Feng Wang, and Dazhi He. Task decoupled framework for reference-based super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5931–5940, 2022. 2

[11] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2103–2112, 2021. 2, 3, 4, 5, 6, 7

[12] Junyong Lee, Myeonghee Lee, Sunghyun Cho, and Seungyong Lee. Reference-based video super-resolution using multi-camera video triplets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[13] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 1, 2, 5, 6, 7, 8

[14] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6368–6377, 2021. 2

[15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 1

[16] Sanghyun Son and Kyoung Mu Lee. SRWarp: Generalized image super-resolution under arbitrary transformation. In *CVPR*, 2021. 2

[17] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10247–10266, 2023. 3, 4, 7

[18] Yifan Wang, Federico Perazzi, Brian McWilliams, Alexander Sorkine-Hornung, Olga Sorkine-Hornung, and Christopher Schroers. A fully progressive approach to single-image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 864–873, 2018. 2, 3

[19] Bin Xia, Yapeng Tian, Yucheng Hang, Wenming Yang, Qingmin Liao, and Jie Zhou. Coarse-to-fine embedded patchmatch and multi-scale dynamic aggregation for reference-based super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2768–2776, 2022. 2

[20] Kai Xu, Ziwei Yu, Xin Wang, Michael Bi Mi, and Angela Yao. Enhancing video super-resolution via implicit resampling-based alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2546–2555, 2024. 2

[21] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5800, 2020. 2

[22] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on*

computer vision and pattern recognition, pages 5728–5739, 2022. 2, 6

[23] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4791–4800, 2021. 2

[24] Lin Zhang, Xin Li, Dongliang He, Fu Li, Yili Wang, and Zhaoxiang Zhang. Rrsr: Reciprocal reference-based image super-resolution with progressive feature alignment and selection. In *European Conference on Computer Vision*, pages 648–664. Springer, 2022. 2

[25] Lin Zhang, Xin Li, Dongliang He, Fu Li, Errui Ding, and Zhaoxiang Zhang. Lmr: A large-scale multi-reference dataset for reference-based super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13118–13127, 2023. 1, 2, 4, 5, 6, 7, 8

[26] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 2, 3

[27] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7982–7991, 2019. 2, 8

[28] Haitian Zheng, Mengqi Ji, Lei Han, Ziwei Xu, Haoqian Wang, Yebin Liu, and Lu Fang. Learning cross-scale correspondence and patch-based synthesis for reference-based super-resolution. In *BMVC*, page 2, 2017. 2

[29] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *Proceedings of the European conference on computer vision (ECCV)*, pages 88–104, 2018. 2